DOCUMENT RESUME

ED 395 951 TM 025 019

AUTHOR Woolley, Kristin K.
TITLE Revised Thinking about the Nature of Score Validity.
PUB DATE 25 Jan 96
NOTE 15p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (New Orleans, LA, January 25, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Models; *Scores; *Standards; *Test Interpretation; *Test Theory; *Validity
IDENTIFIERS *Consequential Evaluation

ABSTRACT

The theory of score validity has undergone several revisions within the measurement community. The current consensus among professionals is a rejection of the trinitarian doctrine (J. P. Guion, 1980) of score validity and the recognition of a unified view that includes social consequences of test interpretation and use. While some aspects of the "unified theory" (S. Messick, 1989) are controversial, the integration of this view is already changing professional standards. This paper reviews the history of score validity and the ongoing research concerning alternate approaches to validity theory. The paper also challenges the attempt to add "consequential validity" as a new validity construct. The concept of social consequences as a validity question leads to a multitude of problems of measurement that researchers may hesitate to address. (Contains 12 references.) (Author/SLD)

REVISED THINKING ABOUT THE NATURE OF SCORE VALIDITY

Kristin K. Woolley

Texas A&M University 77842-4225

Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, LA, January 25, 1996

Abstract

The theory of score validity has undergone several revisions within the measurement community. The current consensus among professionals is a rejection of the trinitarian doctrine (Guion, 1980) of score validity and the recognition of a unified view that includes social consequences of test interpretation and use. While some aspects of the "unified theory" (Messick, 1989) are controversial, the integration of this view is already changing professional standards. This paper reviews the history of score validity and the ongoing research concerning alternate approaches to validity theory. The paper also challenges the attempt to add "consequential validity" as a new validity construct.

Thinking about score validity has evolved throughout the history of psychological measurement. From an early focus on the factor composition of tests (Guilford, 1946) to the more recent emphasis on the influence of score use upon high stakes decisions in our society (Moss, 1995), the concept of score validity has been adapted to the changing contexts of measurement. Given these evolutions, it is not surprising that the concept of validity confuses so many people. The upcoming revision of the 1985 Standards for Educational and Psychological Testing (AERA, APA & NCME, 1985) has propelled score validity once again to the forefront of discussions within the measurement community.

The contemporary consensus among professionals is that score validity is now a more complex and expanding concept with an increasing emphasis on social factors. Some claim (Messick, 1989; Moss, 1995) that test developers hold an indirect yet significant power base that influences social policies. The concept of "consequential validity" (Messick, 1989) has raised the issue of the limits of social responsibility. Others argue (Shepard, 1993) that there may be too much emphasis on the unlimited information that can be gathered concerning social consequences of test use and misuse. The trend taken to the extreme results in a dangerous belief that test developers should control, direct, and implement social policy. But it seems impossible and clearly unscientific to require test developers to take on this responsibility.

Regardless of the direction or role score validity will have in the future, change is inevitable. The definition of score validity is on the verge of further evolution that will effect social, economical, and political events. It is perhaps worthwhile to further examine the evolution of views of score validity to better understand contemporary thought. The present paper will follow the gradual shift from the trinitarian doctrine (Guion, 1980) of score validity toward the new unified view and examine the change that this shift represents for the measurement community. More recent ongoing research concerning alternative approaches to validity theory will also be addressed. The paper will conclude with the controversial topic of consequential validity and its place vis-`a-vis score validity.

Movement Toward a Unified View

According to Guion (1980), consumers of measurement instruments have disregarded or misused score validity because of lack of education, confusion, and frustration. Without experience, many consumers find measurement issues mysterious. These consumers attempt to resolve the existential crisis of their responsibilities by artistically assigning their responsibilities to test developers. Test developers do not want to be burdened with all responsibilities, so as a result score validity is applied in a haphazard and incomplete manner or not considered in the decision factor at all (Guion, 1980).

More recently the trinitarian doctrine of score validity has been rejected. This doctrine identified three separate types of validity: content, criterion-related, and construct. These types of validity were considered interdependent (Shepard, 1993), however, in practice many researchers had already established independent methods of showing score validity. The current consensus is that content and criterion-related validity are two special cases of construct validity. Guion (1980) explains the basic tenets of content and criterion-related validity and how they function as a part of construct validity.

Content validity is defined as the extent to which the instrument's questions are representative of the domain one is trying to measure. To establish proper content validity one must identify the domain and successfully select a sample of items or tasks from this domain. The key point is that the domain (construct) must be identified first before content validity can be evaluated. According to Guion (1980), personnel testing is an example of when construct validity must be established prior to reliance on content validation. Criterion-related validity is used when one is trying to measure how well one variable, like the Scholastic Aptitude Test score (SAT), predicts future performance or college Grade Point Average (GPA). To provide more meaningful results, however, SAT scores and college GPAs must be shown to be a part of the same construct to be measured. Again, criterion-related validity is a part of construct validity. "Both kinds of evidence known as content validity and as criterion-related validity may contribute to how well the operations represent the underlying concept, but they do so

only insofar as they are special cases of construct validity. Construct validity seems to provide the unifying theme" (Guion, 1980, p. 393).

The new view of construct validity or the "unified theory" (Messick, 1989) does not permit professionals to rely on a combined validity coefficient as the answer to the validity question. In the example of the SAT score with college GPA, one must consider the sample and the possibility of criterion contamination where money or political interest may be involved in decisions to admit students. Researchers need to evaluate the reliability of scores on the SAT, college GPA, and how these measures are viewed in different geographic regions. One complication is that the nature of "scholastic achievement" may change over time. It is not sufficient to attach a validity coefficient to a set of scores and claim they are eternally valid.

The concept of a unified theory of validity is still under debate, however, and some specific definitions and responsibilities of researchers have been identified by the Joint Committee for Standards on Education Evaluation (1994). This movement toward a more well-defined concept of score validity promotes sound measurement procedures. The use of this new theory in also protects measurement and mental health professions from legal and ethical challenges. Validity as a measurement concern seems to be receiving considerable attention due to the criticsms of current assessments combined with the growing influence of social, economic, and political concerns related to educational measurement.

Historical Trends

The development of the unified theory is significant when one reviews the literature recording the different conceptualizations of score validity. Guilford (1946) suggested that validity consisted of factorial and practical applications. If an instrument's scores were shown to measure certain factors, then the instrument yielded valid scores. The practical application of validity was when these factors correlated with a different criterion, such as job performance or achievement. Guilford (1946) was also a strong advocate for mathematical and procedural standards that should be employed regularly to ensure objective evaluation of people and performance. He said,

> It is my conviction that only by objective empirical procedure such as factor analysis can we know what abilities and traits are represented in either tests or jobs. It requires such an approach to enable us to break the shackles of tradition and realize the great richness of human variability that actually exists. (Guilford, 1946, p. 433)

Guilford (1946) also suggested that the reliance on IQ was perhaps limiting and that greater reliability and validity in selection of military occupations may be achieved through a series of test batteries with each separate test maximizing factor validity. He predicted that "...any test author will be expected to present information regarding the factor composition of his tests" (Guilford, 1946, p. 438). In the 1950s, other theorists advancedGuilford's emphasis on factor analysis and construct validity.

Cronbach (1989) discussed the evolution of construct validation after 1950 and admitted that the explanation of this classic question is difficult to follow. His definition of construct validation was simple, "...how much to trust the test and in which cases" (Cronbach, 1989, p. 149). Cronbach and Guilford seemed to agree that early tests of general ability did not satisfy the requirement of validity. If a test score measures general ability and this is somehow correlated with job performance then very little information is known. One must be more specific regarding what the test score measures to claim that the score can predict successful job performance or other criteria.

Cronbach (1989) noted that constructs became more focal as more constructs were elaborated by social scientists, and that this made construct validity even harder to establish. In 1959, Campbell and Fiske presented the multitrait-multimethod matrix to further assist researchers in strengthening their validity claims by establishing that scores measure what they are supposed to and only these elements. Although easy to apply and understand, the matrix must be combined with other measures of validity to show the suitability of an instrument scores. Cronbach (1989) explained that score validity is most often questioned when decisions involve possible employment discrimination and test bias. Again, the controversial

definition and application of construct validity was raised and redefined further moving toward a more meaningful analysis of test evaluation.

In Messick's chapter on validity (1989), he followed the progress and changes of score validity throughout the measurement literature. One important aspect he clarified was the unified nature of construct validity and that this view involved both test interpretation and test use. Messick (1989) claimed that a two-facet approach to score validity is required and must involve evidential and consequential levels. The evidential basis of test interpretation and test use includes the unified view of construct validity combined with test relevance and utility. The consequential basis of test interpretation and test use includes value implications and social consequences. Addressing social consequences, he suggested that, "Validity is an inductive summary of both the existing evidence for and the potential consequences of score interpretation and use" (Messick, 1989, p. 13). Not only are the empirical results of test evaluation important in decision making, but the effects of these decisions upon the individuals are also important. The meaningfulness of score interpretation is lost when validity does not include the intended and unintended goals of the test. "The meaning of the measure, and hence its construct validity, must always be pursued--not only to support test interpretation but also to justify test use" (Messick, 1989, p. 17). Messick's chapter (1989) established the significance of the unified theory as well as expanding the view of validity to include value judgments. This idea of validity further challenges and extends the previous system of measurement. Shepard (1993) quoted Cronbach's words,

> We might once have identified validation with a single question, What does the instrument measure? That question will not have an objective universal answer. A yet more judgmental question now takes on equal importance: And why should *that* be measured? (Shepard, 1993, p. 426)

Current Trends

Although defining and establishing construct validation may seem to be never ending process, Cronbach (1989) suggested a set of statements that address one part of the validity

issue. This alternate form of validity inquiry uses an interpretive argument to answer certain questions related to test score validity. According to Kane (1992), the method involves "...explaining the meaning of the test score and, thereby, to make at least some of the implications of the score clear" (Kane, 1992, p. 527). This theory uses logic by presenting a series of arguments that validates assumptions, statements, and decisions. The most important advantage of this approach to validation studies is that the instrument is strengthened by the continual redefinition of the argument. The more information and competing arguments that are interjected, the greater is the chance that the instrument can be adjusted and improved. The result is not a final evaluation of validity, but rather an ongoing process that can be applied and reapplied to any test situation. New demands upon test developers and test consumers have influenced these alternative approaches to score validity. The increased use of high-stakes testing has forced test developers to create more efficient and logical procedures to evaluate these tests.

The upcoming revision of validity theory in the 1985 Standards for Educational and Psychological Testing (AERA, APA & NCME, 1985) is critical to further promoting the understanding and use of validity in research and practice (Moss, 1995). Several issues need to be addressed to properly choose the best perspective on validity theory that fits the current trends in measurement while guarding against oversimplification. Moss (1995) believes the purpose of the revision is not only to reflect these trends, but to shape future policy and practices within the profession. Researchers and consumers must have guidance as to

> what evidence is necessary to justify the use of an assessment from what evidence is the ongoing responsibility of the measurement community at large--evidence desirable to enhance theory and practice in the long run, but beyond what can be reasonably expected of a particular developer or user. (Moss, 1995, p. 7)

The issues that require attention are present both within the traditional concepts of validity as well as in new ideas not yet well established or accepted.

Moss (1995) recommended several changes to the 1985 *Standards*. Her first recommendation was that the definition of validity go beyond the currently accepted unified view of construct validity. She said, "The essential purpose of construct validity is to justify a particular interpretation of a test score by explaining the behavior which the test score summarizes"(Moss, 1995, p. 6). Under the definition of score validity such concepts as convergent and discriminant evidence as well as the requirement of competing arguments to construct validity should be included. Moss (1995) also suggested that alternatives to the three categories of validity evidence be provided in the revision. The main reason for this is that not all research requires the same types of validity evidence. Validity categories could be expanded to include questions concerning test use justification. The redefinition of score validity with the new version of the *Standards* will affect all sides of the measurement community. It is necessary for professionals to be cognizant of the direction these revisions are taking.

More recent thinking in the area of score validity dealt with performance testing and some concerns about the gap between validity theory and practice. Moss (1992) emphasized the need for researchers to expand their models of validity inquiry, especially in the area of performance assessment. It is common that alternate methods of testing are utilized in the area of high-stakes decisions (i.e., oral examinations for the Ph.D.). More sophisticated research is necessary to adequately evaluate the validity of scores on these non-standardized tests. Administrators, teachers, and students base curriculum, instruction, and learning decisions on how they will be tested. There are several methods of validity inquiry that appropriately address performance issues and balance the technical and consequential considerations of test use.

> All of the authors ground their analysis of the concept of validity in an argument about consequences: the consequences of performance assessment are likely to be more beneficial to teaching and learning than are the consequences of multiple-choice assessment alone. (Moss, 1992, p. 248)

It seems that the people who make decisions about which type of validity procedures are used for which tests have a great deal of power. Moss (1992) suggested that the measurement profession possesses this power to influence the scope of performance testing.

Other concerns about the use of score validity in high-stakes situations were identified by Shepard (1993). She discussed individual cases that show how score validity and score application can often distort decisions. An example of how score validity studies are often one-sided is the use of the SAT in college admissions decisions. The logic of this practice is based upon the correlation between test scores and first-year college grades. It is well known, however, that correlations alone do not describe a relationship. Although certain academic attributes related to college success have been shown to influence the SAT score, other factors such as high school curriculum and testing courses may also affect scores. Shepard (1993) suggested that for women, success in college is related to motivation not measured by the SAT. Explorations of the alternatives to using the SAT are underway. The use of the General Aptitude Battery (GATB) for referral decisions in employment is another example where score validity is part of the gap between research and practice (Shepard, 1993). An extensive review of the usefulness of this instrument was conducted and results indicated that several problems exist that threaten score validity within the development and administration of the test. Because (a) the norms used to check the applicability of the GATB across all jobs are outdated and (b) the test is timed, the test may be biased against slower responding employees in job areas that do not have time limits (Shepard, 1993). According to the GATB review, the test has major flaws that directly affect score validity. A more salient consideration may be the consequences involved in employment decision errors due to improper measurement procedures.

Both Moss (1992) and Shepard (1993) agreed that critical decisions based upon test results require test developers and test consumers to further challenge oversimplified answers to score validity questions. A primary component of reviewing the validity question is the ability to show evidence of competing arguments and that testing is necessary in the first place. Consequences of test use and interpretation are seen as necessary investigations; however, it is

unclear how far beyond the boundaries of validity research test developers are expected to venture.

Consequential Validity

An issue presently facing the measurement community concerns the concept of social consequences as a validity question. Consequences of test interpretation and use include both the actual goals of the instrument and the unintended outcomes (Messick, 1989). The 1985 Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1985) indicated researchers must attend to the "...appropriateness, meaningfulness, and usefulness of specific inferences made from test scores" (p. 9). Most agree (Kane, 1992; Shepard, 1993) that it is important to show the test is useful for a particular purpose, checking its immediate consequential validity. It is more difficult, however, to identify all the other possible outcomes that the test may influence in addition to its intended function. Moss (1995) argued that including consequential validity in the revised *Standards* does not prescribe social policy, but requires researchers to provide pertinent information so that social decisions can be made. It may be unreasonable, however, to guard against bias in this social "information" without a clear separation between researchers and consumers.

The social, economic, and political implications of creating a new construct called consequential validity are numerous. Consequential validity is achieved by involving primary users in test evaluation to enhance the utilization of results. Brandon, Lindberg, and Wang (1993) describe how consequential validity can be beneficial in reviewing educational programs. The problem, however, is choosing the primary users to involve. Since these groups are not static, it would seem that consequential validity studies would have to be repeated infinitely in response to changing groups and intended test use. The resources required would be excessive. Other requirements to effectively measure consequential validity are that involved groups remain small and that each party can communicate opinions. These criteria are difficult to meet for instance when addressing an evaluation of kindergarten children's attitudes and teaching methods. Another issue is that primary users are not trained in

psychometric procedures. Reliance upon consequential validity may confound scientific views of score integrity. A more serious issue facing test developers is that by attempting to measure consequential validity developers become involved in the decision-making process through their influence. Scientists may find themselves in the middle of power disputes between stakeholders and beneficiaries. It seems that the measurement profession would be better off without the responsibility of consequential validity.

The changing trends in validity theory over time have shown that this area of measurement is dynamic and continues to be redefined. Several important theories have improved measurement and have encouraged a greater awareness of the effects of score validity. The measurement community, however, cannot be expected to control all of the variables in test misinterpretation and misuse even in a society that is largely testing focused. It would be unfair to assert that all equality and social progress depends upon test manufacturers and their procedures. Although high-stakes decisions are regularly based upon the results of test evaluations, the primary job of creating and implementing educational policies must remain outside the duties and responsibilities of the measurement community.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement and Education. (1985). Standards for educational and psychological testing. Washington, DC: Author.

Brandon, P. R., Lindberg, M. A., & Wang, Z. (1993). Involving program beneficiaries in the early stages of evaluation: Issues of consequential validity and influence. Educational Evaluation and Policy Analysis, 15, 420-428.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), Intelligence: Measurement, theory and public policy (pp. 147-171). Chicago: University of Illinois Press.

Guilford, J. P. (1946). New standards for test evaluation. Educational and Psychological Measurement, 6, 427-439.

Guion, R. M. (1980). On trinitarian doctrines of validity. Professional Psychology, 11, 385-398.

Joint Committee for Standards on Educational Evaluation. (1994). The program evaluation standards: How to assess evaluations of educational programs (2nd ed.). Newbury Park, CA: Sage.

Kane, M. T. (1992). An argument-based approach to validity. Psychological Bulletin, 112, 527-535.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment. American Psychologist, 50, 741-749.

Moss, P.A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. Review of Educational Research, 62, 229-258.

Moss, P. A. (1995). Themes and variations in validity theory. Educational Measurement: Issues and Practice, Summer, 5-13.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), Review of research in education (Vol. 19, pp. 405-450). Washington, DC: American Educational Research Association.